Encevra Research Round 50

Title: When Both Readers Are Wrong: Open Removal Tests for Review and Rewriting

Round code: ORT-50

Creator: Wilkins, Benjamin. Encevra. ORCID: 0009-0005-6532-3200

Date: November 7, 2026

Status: All experimental phases attempted. Five individual experimental outputs were empty or incomplete and were not repeated. Prepared for review, not published by this run. DOI pending reservation.

Abstract

The Theory of Collaborative Intelligence roadmap calls Round 50 an open removal test: joint error when both roles can fail in the same direction, and rewrite gate behavior on genuinely wrong rewrites. Round 49 supplied two unresolved observations: Witnesses C and D shared a confident FAIL on near miss items, while a stored procedure helped only the reviewer to which it was given. It also left the gate's live catch rate unmeasured and identified questionable whole-instrument exhibit labels. Round 50 removes the second read and the stored procedure on fresh contractual and non-contract policy packets, and removes a status-preservation gate on live ledger rewrites from both correct and deliberately wrong starting ledgers.

Twenty-four authored packets cover three contractual and three non-contract families. Six new SEC exhibit windows ask a deliberately narrower receipt or delivery question, not whether an entire agreement is effective. Witness A agreed with all 30 author labels but flagged a mismatch between specific required instrument names and generic approval records. Witness B wrote procedures from separate training vignettes. Witnesses C and D each read every authored packet with and without procedure memory. Witness C rewrote both a correct and a wrong ledger for each authored packet and exhibit window.

Structurally, an exact quote gate admits all 48 wrong status mutations, while status preservation rejects all 48. Four of 27 truth and paired-verdict states remain unsafe despite agreement routing. Empirically, the readers produced no decisive unsafe verdict on usable authored reads, so the planned same-wrong decisive failure did not materialize. They did share errors in the other direction: five jointly wrong pairs with no procedure, four with procedures, all false holds. On the 21 items with all four usable reads, paired endpoint accuracy is exactly 16 of 21 under both arms. Seven of 29 usable rewrites from correct inputs became wrong false holds while retaining valid quotes. Status preservation routed all seven but did not improve their endpoint outcome, because routing also means HOLD. From wrong inputs, the validator made 22 correct repairs in 30 outputs; preservation routed all 22 repairs, left eight wrong unchanged statuses admissible, and reduced correct final outcomes from 22 to 12 of 30. The gate protects a trusted starting state, not truth itself. A gate that prevents damage can also prevent recovery.

1. Archive check and continuity

The latest available project entry before this run is Round 49, WMR-49. No Round 50 manuscript, route or registry entry existed. The accessible archive was checked through Round 49. Six early numbers, 1, 2, 4, 5, 6 and 7, have neither registry entries nor manuscripts. Rounds 3, 8, 9 and 10 have summary entries but no full manuscripts; Round 15 is an internal notebook. Round 24 has a manuscript but no linked round page. Consequently this is a strict comparison against all available history, not an assurance about missing material.

Round 35 studied unanimous falsehood, Round 42 measured joint misreadings across roles and witnesses, Round 43 constructed span-valid wrong rewrites, and Round 44 restricted validator releases by the original decision. These are antecedents, not new discoveries claimed here. Round 50's new comparison is removal on fresh mixed-domain outputs and a live correct-versus-wrong starting-ledger comparison that exposes the preservation gate's recovery cost. Procedures are now given to both reviewers and derived from a separate training vignette

rather than an evaluated positive packet. All six exhibit accessions are new against the recorded accession set through Round 49 and were fetched again in this run. Details are in the archive audit and exhibit record.

The exact roadmap question is: "Open removal tests, not another contract refinement: joint error when both roles can fail in the same direction, and rewrite gate behavior on genuinely wrong rewrites."

2. Fixed plan and hypotheses

The plan was saved before live calls. E50-A is deterministic. E50-B contains the live blind review, procedure generation and independent reads. E50-C contains live ledger rewrites. A final synthesis is commentary, never gold. Findings, interpretations and structural demonstrations remain separate.

ID	Fixed statement
H50.1	At least one same-wrong decisive agreement survives routing in each domain under no memory.
H50.2	Procedure memory reduces same-wrong decisive agreements without more false holds in either domain.
H50.3	Exact-span validation admits at least one naturally wrong rewrite from a correct input.
H50.4	Status preservation routes every naturally wrong status change from correct inputs, and every status-changing repair of a wrong input.
H50.5	Removing the second read does not improve safety, though it can improve endpoint accuracy by avoiding false holds.
H50.6	Structurally, quote validity cannot certify semantic status and agreement cannot certify truth.

These are output-replay ablations, not a full live factorial orchestration experiment. They do not compare equal numbers of calls; the fair resource comparison remains Round 52. No run result was used to select or rewrite a packet after execution began.

3. Materials and labels

Domain A contains escrow return authorization, franchise territory activation and equipment lien release. Domain B contains municipal water advisory rescission, observatory dataset embargo clearance and museum pest quarantine clearance. Domain B consists of non-contract policy records and is not SEC material. All 24 are authored analytic examples, not actual professional instructions or evidence that the listed real-world actions occurred.

Every family has one authorization rule and four variants:

• P: an approval signed by the designated role. Author label PASS.

• N: an approval signed by an explicitly unqualified person, with an open file. Author label HOLD, because a separate qualifying record is neither proved nor excluded.

• F: the N record plus an exhaustive attestation that no other qualifying record exists at the audit time. Author label FAIL.

• R: the N record plus a separate qualifying approval. Author label PASS.

The reviewer instructions explicitly distinguish a defective listed record from proof that no other qualifying record exists. This makes the baseline stronger and simpler than Round 49's aggregate requirement. It limits the attempt to reproduce Round 49's unsafe near miss error: the new materials are not a matched hard-pool replication of that

result.

Witness A reviewed all 24 packets and six exhibit windows blind to the author labels at maximum reasoning and agreed with all 30. It reported a drafting defect on positive and cure variants: the rule requests a named instrument, such as a release instruction or clearance, while the record uses the generic word approval. It accepted the intended signer-based interpretation but noted that the wording is not perfectly aligned. The labels were not silently changed. Several later reviewers treated this mismatch as grounds for HOLD. Those outcomes are errors relative to the stated author labels, not independently settled professional ground truth.

The six SEC windows isolate one named requirement for delivery or receipt. All six state a condition but do not establish that it actually happened within the selected window, so the window label is HOLD. This does not mean the agreement is ineffective or that delivery never occurred outside the window. It deliberately avoids the questionable whole-instrument CERTIFY rule identified in Round 49.

Item	Filer	Exhibit	Filing date	Accession
Z01	MariaDB plc	EX-10.1	2024-01-11	0001929589-24-000004
Z02	RING ENERGY, INC.	EX-10.1	2024-02-16	0001628280-24-005338
Z03	FTAI Aviation Ltd.	EX-10.11	2023-04-27	0001590364-23-000013
Z04	Telesis Bio Inc.	EX-10.1	2023-08-11	0000950170-23-041561
Z05	TELA Bio, Inc.	EX-10.25	2024-03-22	0001558370-24-003782
Z06	American Acquisition Opportunity Inc.	EX-2.1	2023-05-01	0001654954-23-005446

The exhibit record includes the exact source URLs, audit windows and target spans. Source normalization collapses whitespace only. Target spans are checked against freshly fetched text. A narrow window question cannot support an inference about the entire filing.

4. Readers, procedures and removal rules

Witness B received one separate training vignette per family. Each vignette describes a defective signer and an exhaustive negative register, with its solved FAIL label. It was asked for a reusable method that also handles an open file and a later qualifying approval, without carrying a verdict or quotations into the procedure. No evaluated packet supplied a card. This removes the specific positive-source overlap of Round 49, but the training instructions still reveal the family structure, so this is not independent skill discovery.

M0 means packet only. MP adds the stored family procedure above the packet. Witnesses C and D receive fresh calls and never see each other's verdicts. Separate calls are not proof of statistically independent errors. A paired outcome releases PASS or FAIL only when both usable verdicts agree on that decisive value; otherwise it is HOLD. Removing the second read retains the corresponding single verdict on the same usable pair set.

A verdict is unsafe when it is PASS or FAIL and differs from gold. A false hold is HOLD on gold PASS or FAIL. Both readers wrong includes shared false holds as well as decisive errors. Same-wrong decisive agreement is the narrower unsafe agreement that a disagreement router cannot catch. These endpoints must not be conflated.

Missing or incomplete JSON is not converted into an empirical HOLD. It is recorded as unavailable and excluded from complete-case endpoint denominators. In a deployed policy it would require review, but no operational completion outcome is invented here. Comparisons between memory arms use the 21 items for which both readers produced usable outputs under both arms.

5. Rewrite gates and deterministic checks

A ledger has a status and an exact packet quote. SATISFIED maps to PASS, NOT ESTABLISHED to HOLD, and REFUTED to FAIL. Gspan checks the status vocabulary and nonempty quote membership. It does not check whether the quote entails the status. Gdelta adds equality of the proposed status to the input ledger's status. It has no access to gold. If a valid proposal is not admitted, the routed endpoint is HOLD. It does not silently restore the original verdict.

Witness C receives both a correct input ledger and a deliberately wrong one for every authored packet and exhibit window. Wrong seeds cycle SATISFIED to NOT ESTABLISHED, NOT ESTABLISHED to REFUTED, and REFUTED to SATISFIED, without changing the quote. Inputs are interleaved by a fixed item shuffle. A genuinely wrong live rewrite from a correct input is a usable returned status that differs from the author gold; it is not the deliberately seeded input and not a scripted mutation.

Structural findings

• **F50.S1:** All 48 wrong status mutations retain valid quotes and pass Gspan. Gdelta rejects all 48 and admits all 24 unchanged correct ledgers. This is a finite design result, not a measured live safety rate.

• **F50.S2:** Across all 27 combinations of three truth values and two three-valued reader outputs, agreement routing leaves four unsafe states. Both readers can agree on the same wrong decisive output. Separate role names do not remove those states.

• **F50.S3:** Preserving a starting status necessarily rejects a repair that changes that status, even when the repair is right. The gate cannot distinguish improvement from damage without an additional trusted semantic adjudication step.

These restate the boundary exposed by prior composition and rewrite studies while testing a new recovery condition. They do not establish that a proposed universal semantic gate exists.

6. Empirical reader results

Arm	Reader	Correct	Unsafe	False holds	Usable of 24
M0	Witness C	16	0	7	23
M0	Witness D	18	0	5	23
MP	Witness C	17	0	7	24
MP	Witness D	18	0	4	22

The M0 calls for A01R by Witness C and B03P by Witness D returned truncated JSON. MP calls for A02P and A01R by Witness D returned empty responses. The corresponding decisions are unavailable. There were no transport errors or budget-stopped calls.

Arm	Domain	Usable pairs	Both wrong, any verdict	Same wrong decisive agreement	Disagreements
M0	A	11	3	0	1
M0	B	11	2	0	0
M0	all	22	5	0	1
MP	A	10	3	0	0
MP	B	12	1	0	2
MP	all	22	4	0	2

No same-wrong decisive agreement was observed in either arm or domain. H50.1 is not supported. The attempt did observe jointly wrong answers in the same conservative direction: all both-wrong pairs were false holds. These can preserve safety while losing useful authorizations; they are not unsafe releases.

Raw arm denominators differ. The matched memory comparison therefore uses the same 21 items under both arms:

Arm	Domain	Common items	Correct paired outcome	False holds	Both readers wrong
M0	A	10	7	3	2
M0	B	11	9	2	2
M0	all	21	16	5	4
MP	A	10	7	3	3
MP	B	11	9	2	1
MP	all	21	16	5	4

Procedure removal changes neither paired accuracy nor paired false holds on these common items. Domain A gains one shared error while Domain B loses one. H50.2 is not supported as a joint-error reduction claim. It is not evidence that procedures never help: the baseline already explains the open-record distinction, the pool is small, and some labels contain the acknowledged wording mismatch.

On each arm's complete-pair set, removing the second read and retaining Witness D raises correct outcomes from 16 to 17 of 22 under M0 and from 16 to 18 of 22 under MP, with zero unsafe verdicts in every comparison. Retaining Witness C gives 16 of 22 in both arms. Thus the second read buys no observed safety improvement here and can add false holds. H50.5 is consistent with the observations but is not a demonstrated safety benefit; its no-improvement safety direction is also a property of the conservative replay rule.

7. Empirical rewrite results

Starting ledger	Domain	Usable	Wrong live outputs	Quote-valid wrong outputs	Correct status repairs	Repairs routed by Gdelta
correct	A	11	4	4	0	0
correct	B	12	3	3	0	0
correct	SEC	6	0	0	0	0
correct	all	29	7	7	0	0
wrong	A	12	5	5	7	7
wrong	B	12	3	3	9	9
wrong	SEC	6	0	0	6	6
wrong	all	30	8	8	22	22

One correct-input rewrite, A01R, returned truncated JSON. Its status is not scored as a natural error. The 29 other correct-input rewrites contain seven status changes that are wrong relative to author gold, all from decisive truth to HOLD. Every one carries exact packet words. This supports H50.3 as written, but not the stronger claim that a live unsafe decisive rewrite occurred. The seven are A01P, A02P, A03P, A03R, B01P, B02P and B03P. Their recurring rationale is that a generic approval is not necessarily the specifically named instrument, the same drafting issue the blind review flagged.

Gdelta routes all seven natural status changes from correct inputs. Since the changed verdict and the routing verdict are both HOLD, this makes no difference to endpoint accuracy: raw, Gspan and Gdelta are all correct on 22 of 29 usable outputs, with seven false holds and no unsafe verdict. Detecting a change is not the same as repairing its effect.

From wrong inputs, Witness C produces 22 correct status-changing repairs out of 30. Gspan admits them. Gdelta routes all 22, as predicted, while admitting eight wrong unchanged statuses. Those eight are false holds, not decisive unsafe outputs. Endpoint correctness falls from 22 of 30 behind Gspan to 12 of 30 behind Gdelta; false holds rise from eight to eighteen. Twelve of the repairs end at a correct HOLD and remain correct after routing; ten correct decisive repairs are lost. All six exhibit repairs are to HOLD, so routing them has no endpoint penalty on the limited window question.

H50.4 is supported as a behavior statement about this deliberately restrictive gate: it routes seven natural damaging changes and every one of 22 repairs. That is not support for deploying it as a truth filter. Removing Gdelta improves measured utility on wrong baselines without an observed safety loss. Baseline quality is part of the gate's safety contract.

8. Findings versus interpretations

Empirical findings

• **F50.E1:** Blind review agrees with 30 of 30 labels but flags named-instrument versus generic-approval wording on positive and cure variants.

• **F50.E2:** Decisive unsafe reads and same-wrong decisive agreements are zero on all usable authored reads and pairs. Joint false holds remain: five M0 pairs and four MP pairs on their respective complete-pair sets.

• **F50.E3:** On the common 21-item memory comparison, both paired arms are correct on 16 with five false holds. Both-reader errors total four under both arms, with opposite one-item changes across domains.

• **F50.E4:** Seven of 29 usable correct-input rewrites become wrong HOLD statuses with exact quotes. Gdelta routes all seven but does not improve the HOLD endpoint.

• **F50.E5:** Wrong-input rewriting repairs 22 of 30 ledgers. Preservation routes all 22 repairs, admits eight wrong unchanged statuses and reduces final correctness from 22 to 12 of 30.

• **F50.E6:** Five individual outputs are unusable: two empty MP responses, two truncated M0 responses and one truncated correct-input rewrite. None was repeated or silently converted into a scored HOLD.

Interpretation

The broad claim that a second read or a preservation gate improves a collaboration does not survive as an unconditional statement. Their contribution depends on error direction and on what state is trusted. In this pool, both readers can be wrong together by withholding rather than by releasing. The second read does not improve safety on zero-unsafe input streams and sometimes loses a correct decisive answer. The preservation gate protects the initial status, including when that status is wrong. It can detect a harmful status change while producing the very same harmful HOLD as its fallback.

The strongest surviving domain-general statement is therefore conditional: a preservation mechanism needs an explicit trust boundary for the state it preserves and a separately tested recovery path. Quote identity and unchanged status are trace-integrity checks, not semantic truth checks. Agreement is a fact about the readers, not independent evidence that their conclusion is true. Rounds 51 and 54 must test these limits on tasks that are not another document authorization packet.

9. Hypothesis outcomes

ID	Outcome	Reason
H50.1	Not supported	No same-wrong decisive agreement in either domain.
H50.2	Not supported	No unsafe agreement reduction to measure; common-item paired accuracy and false holds are unchanged.
H50.3	Supported narrowly	Seven quote-valid wrong live status changes, all conservative false holds on potentially contestable wording.
H50.4	Supported as gate behavior	Seven damaging changes and 22 correct repairs routed; eight wrong unchanged statuses admitted.
H50.5	Consistent, safety benefit unmeasured	Removing the second read can recover one or two correct outcomes; all compared unsafe counts are zero.
H50.6	Structurally supported	Forty-eight semantic mutations pass quote checking; four joint-verdict states remain unsafe.

10. Synthesis and analysis audit

Witness A reviewed the completed analysis at maximum reasoning. The first synthesis saw provisional scoring, including the malformed correct-input rewrite counted as a wrong HOLD, and its own output stopped before completing JSON. That commentary is preserved verbatim but is not the final synthesis. A separate concise final audit received the corrected exclusions, common-item comparison and explicit finite-state derivations. No experimental response was repeated.

The final audit agrees that the structural counts are by construction, not live rates; no unsafe-prevention benefit was measured; pairing never beats the better single reader here; and the preservation gate's wrong-baseline utility loss is the relevant live result. It emphasizes that author gold and the acknowledged instrument-name mismatch make false-hold counts interpretation-dependent, that cyclic seeding is not randomized, and that missing outputs cluster on positive cases. If the contested HOLD answers are accepted under a stricter instrument-name reading, the claimed natural rewrite damage weakens or disappears. The seven count is therefore a scoring result, not an independently settled semantic fact.

Its call count of 159 describes the ledger it received before its own final call; the final ledger contains 160 calls, including 158 planned experimental calls and two synthesis commentaries. The record contains 158 nonempty responses, three of which have truncated experimental JSON, plus one truncated provisional commentary; transport status alone is not schema validity. There are 153 usable experimental responses and one usable final commentary. The final audit proposed rerunning missing calls and re-adjudicating wording as future checks. Neither was done in this run. All commentary remains subordinate to the actual transcript and replayed counts.

11. Completion disclosure and limitations

All planned experimental phases were attempted. Five individual experimental responses were empty or cut off and were not repeated. Those decisions are unavailable, not verdicts. The analysis was corrected before the manuscript tables were frozen to distinguish malformed output from a wrong live rewrite and to compare memory arms on common items. The original prompts and responses remain verbatim in the transcript. No execution cap was increased.

The same rule template appears in both domains, so this is mixed material but a narrow task transfer. It does not establish transfer to arbitrary scientific, numerical or software tasks. The baseline explicitly teaches the crucial open-record distinction and is simpler than Round 49's aggregate task. Consequently the zero unsafe result does not refute the earlier shared unsafe error. The named-instrument mismatch makes some false holds defensible, despite blind agreement with the intended labels. There is no external human adjudication. Separate calls do not ensure independent errors. The finite enumeration is not an empirical event frequency. These are deterministic replays over live outputs, not a fair matched-call architecture comparison or full factorial live deployment. Neither the second read nor either gate has a measured unsafe-prevention benefit on this pool.

12. Next checks for the theory program

1. Round 51 should use its planned clearly different task, such as code review against a specification or numerical claims against raw tables. Carry forward baseline trust and recovery as testable conditions, not contract terminology.

2. Remove the named-instrument versus generic-approval ambiguity on a fresh, held-out set, while separately retaining genuinely unresolved cases. Do not retroactively relabel this run.

3. Test recovery that routes changed statuses to a genuinely independent adjudicator with access to raw evidence. Do not define semantic correctness by the starting ledger.

4. Replicate the Round 49 hard near misses directly before interpreting the absence of unsafe joint error as a change in capability. Keep that replication separate from this new-family removal test.

5. Round 52 must compare equal-call baselines and account for review workload and correct decisive repairs lost, not only unsafe releases prevented.

13. Reproducibility

The archive includes the fixed plan, continuity audit, material builder, all 24 packets, exact SEC windows and links, source verification, deterministic state enumeration, live runner, verbatim transcripts, independent blind review and synthesis, replay analysis and verifier, manuscript and PDF builder. The private execution configuration and accounting are not required to replay the recorded evidence and are not included. The verifier makes no live calls. Structured packet fields and complete-case exclusions make each reported count recomputable. The research webpage, PDF and archive remain staged for approval, with no DOI claim until a DOI is reserved.

14. Verified references

• [S98] John C. Knight and Nancy G. Leveson. An experimental evaluation of the assumption of independence in multiversion programming. IEEE Transactions on Software Engineering SE-12(1), 96-109, 1986. doi:10.1109/TSE.1986.6312924. The publisher-indexed title, authors, pages and abstract were verified; a direct DOI fetch timed out. The work concerns dependent program failures and motivates, but does not validate, these reader rates.

• [S72] Zana Bucinca, Maja Barbara Malaya and Krzysztof Z. Gajos. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. Proceedings of the ACM on Human-Computer Interaction 5, CSCW1, 188, 2021. doi:10.1145/3449287. https://arxiv.org/abs/2102.09692. The full paper was opened; its human-study scope and usability tradeoff limit direct transfer to automated reviewers.

• [S40] Matt Gardner et al. Evaluating Models' Local Decision Boundaries via Contrast Sets. 2020. https://arxiv.org/abs/2004.02709. The full paper was opened, including the warning that contrast sets have negative rather than universal positive predictive power.

• [S30] U.S. Securities and Exchange Commission. EDGAR Archives. https://www.sec.gov/Archives/edgar/data/. Six exact source exhibits were opened and verified for this run; the source record supplies each full URL and

accession.